SDCI Net: Collaborative Research: An integrated study of datacenter networking and 100 GigE wide-area networking in support of distributed scientific computing

Zhengyang Liu
University of Virginia

Oct 29, 2012

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

GridFTP Usage Log Analysis

- Purpose: characterize wide area data movement
- Obtained logs from
  - NERSC-ORNL (Sept. 2010)
  - NERSC-ANL (Mar. 4 - Apr. 22, 2012)
  - NCAR-NICS (2009-2011)
  - SLAC-BNL (Feb. 10 - Apr. 26, 2012)
- Analysis performed
  - Throughput Variance Analysis
  - Session Analysis

* Logs obtained with help from John Dennis, NCAR, Ian Foster and Linda Walker, ANL, Jason Hick and Jason Lee, NERSC, Yee-Ting Li and Wei Yang, SLAC, Soctt Bradley and John Bigrow, BNL, Galen Shipman and Scott Atchley, ORNL, and Victor Hazelwood, NICS

GridFTP Usage Log Analysis

```
1 DATE=20121014120031.536839 HOST=dtn02.nersc.gov PROG=globus-gridftp-server
   NL.EVNT=FTP_INFO START=20121014120003.275074 USER=zl4ef FILE=/dev/zero BU
  FFER=1048576 BLOCK=262144 NBYTES=21494235136 VOLUME=/ STREAMS=1 STRIPES=1
  DEST=[134.79.198.146] TYPE=RETR CODE=226
```

- The usage log provides useful information such as:
  - Number of bytes transferred
  - Transfer start date
  - Transfer end date
  - Source/Destination
  - Configuration (streams / stripes settings)

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

Session Analysis

- Question: in spite of increasing rates (which means shorter transfer durations), are transfer sizes large enough to justify VC setup overhead?
  - Our initial finding: transfers are short lived
  - However, as John Dennis pointed out: scientists transfer files, large or small, in batches
  - These transfer "sessions" may be long lived

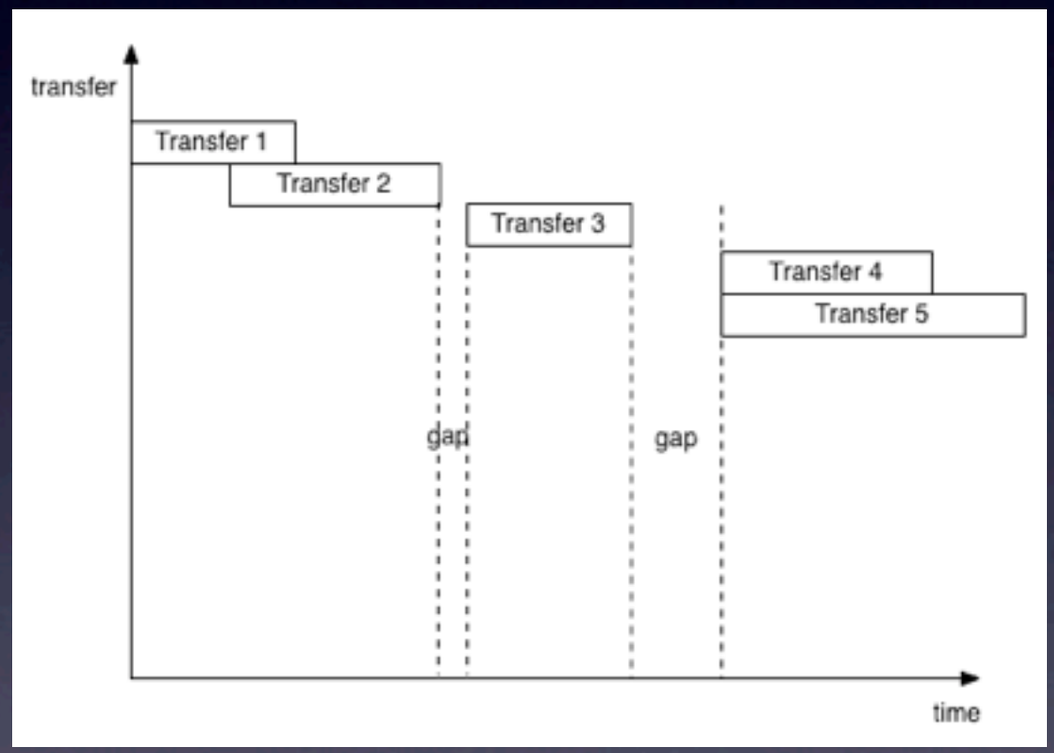
copy2frost

```
1  #!/usr/bin/tcsh -v
2
3  setenv LID "`date +%y%m%d-%H%M%S`"
4  set yrlist = "0120 0121"
5  set CASE = "LRIE03"
6  echo 'file transfer log' >& logs/copy2frost.log.$LID
7  foreach yr (${yrlist})
8     set dirlist = "atm lnd cpl ice ocn"
9  #    set dirlist = "atm lnd cpl ice"
10 #    set dirlist = "lnd cpl ice ocn"
11 #    set dirlist = "ocn"
12 #    set dirlist = "ice"
13    foreach dir (${dirlist})
14         date >> logs/copy2frost.log.$LID
15     find ${dir} -name "*${yr}*" -exec
   globus-url-copy -vb -tcp-bs 2048KB -stripe -p 36
   gsiftp://gridftp.nics.utk.edu/lustre/scratch/jdennis
   /archive/${CASE}/'{}'
   gsiftp://frost.ncar.teragrid.org/ptmp/dennis/${CASE}
   /'{}'  \; >> logs/copy2frost.log.$LID
16
17    end
18 end
```

Line: 1 Column: 1 Shell Scri... Tab Size: 4

Session Analysis

- Applies to transfers between the same two hosts. (i.e. same IP routed path)
- Batch transfers can be "merged" into sessions for analysis purposes
- Transfers in a session:
  - overlap each other
  - small gap (denoted as *g*) between back-to-back transfers



Session Analysis

- Long lived (>10min, defined as 10x the VC setup delay used in ESnet) sessions would justify the use of VC
- However, durations may shorten under high throughput (if VC improves throughput)
- Hypothetical study:
  - What percentage of sessions would have lasted long enough (>10x setup delay) if throughput is assumed to be as high as the third-quartile across all transfers?

Session Analysis

- Pick g = 1 min, the current VC setup delay

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

Session Analysis

- Pick g = 1 min, the current VC setup delay

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

Session Analysis

- Pick g = 1 min, the current VC setup delay

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

Session Analysis

- Pick g = 1 min, the current VC setup delay

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

- Largest session in SLAC-BNL: ~12TB @ 1.06Gbps
- Longest Session in NCAR-NICS: 2.4TB, 13 hrs and 24 mins, 410Mbps

Session Analysis

- Pick g = 1min, the current VC setup delay
- NCAR - NICS
  - 52,454 transfers -> 211 sessions
  - 56.87% sessions (90.54% transfers) would have lasted > 10mins if they have high throughput (defined as 3rd quartile)
- SLAC - BNL
  - 1,021,999 transfers -> 10,199 sessions
  - only 12.5% sessions would have lasted > 10mins
    - however, these 12.5% sessions came from 78.38% of all transfers

TABLE IV: Percentage of sessions suitable for using VCs (percentage of transfers)

	NCAR data set		SLAC data set	
setup delay	1 min	50 ms	1 min	50 ms
$g = 0$	0.12% (2.14%)	87.09% (89.33%)	1.95% (39.41%)	52.58% (89.68%)
$g = 1$ min	56.87% (90.54%)	92.89% (98.04%)	12.54% (78.38%)	93.56% (99.73%)
$g = 2$ min	62.16% (90.71%)	94.59% (98.17%)	15.93% (85.49%)	94.47% (99.85%)

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

Throughput Characterization

- Question: are science data transfer rates high or still low?
- Max. throughput:
  - NERSC-ORNL: 3.64 Gbps
  - ANL-NERSC: 2.76 Gbps
  - NCAR-NICS: 4.23 Gbps
  - SLAC-BNL: 2.56 Gbps
- Significant fractions of link capacity (10 Gbps)

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

Throughput Variance

- Significant variance observed
  - NCAR-NICS: CV = 63.05%
  - SLAC-BNL: CV = 115.24%

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

Throughput Variance

- Significant variance observed
  - NCAR-NICS: CV = 63.05%
  - SLAC-BNL: CV = 115.24%

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

Throughput Variance

- Significant variance observed
  - NCAR-NICS: CV = 63.05%
  - SLAC-BNL: CV = 115.24%

TABLE I: NCAR-NICS sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8,793 (bytes)	5,808.7	70,708.4	263,771.4	320,600	2,873,868.5

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	210.5	1,445	4,039	5,261	48,420

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.1 (bps)	298	468.3	506.1	682.2	4,227

TABLE II: SLAC-BNL sessions and transfers; $g = 1$ min

Characterization of session sizes, in MB

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
812 (bytes)	273	1,195	24,045	4,860	12,037,604

Characterization of session durations, in s

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	18.95	58.92	315.9	151.1	95,080

Characterization of transfer throughput, in Mbps

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.003	22.57	112.8	130.4	183.1	2,560

Throughput Variance

- Significant variance observed
  - NCAR-NICS: CV = 63.05%
  - SLAC-BNL: CV = 115.24%
- We identified a subset of 145 32GB transfers between NERSC and ORNL
- Even though these transfers occurred across the same path, there was considerable variance

TABLE V: The 32GB NERSC-ORNL transfers (145)

	Duration (s)	Throughput (Mbps)
Min	75.4	757.9
1st Qu.	141.2	1251
Median	183.4	1499
Mean	186.6	1625
3rd Qu.	219.7	1947
Max	362.7	3644

Throughput Variance

- Significant variance observed
  - NCAR-NICS: CV = 63.05%
  - SLAC-BNL: CV = 115.24%
- We identified a subset of 145 32GB transfers between NERSC and ORNL
- Even though these transfers occurred across the same path, there was considerable variance

TABLE V: The 32GB NERSC-ORNL transfers (145)

	Duration (s)	Throughput (Mbps)
Min	75.4	757.9
1st Qu.	141.2	1251
Median	183.4	1499
Mean	186.6	1625
3rd Qu.	219.7	1947
Max	362.7	3644

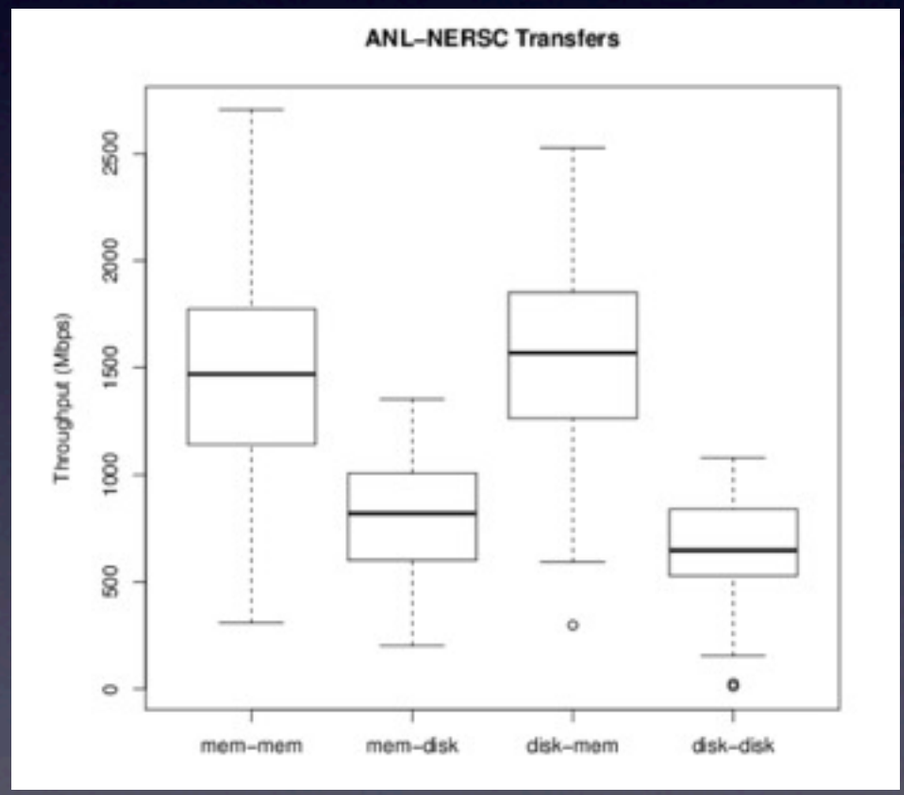
CV = 32.58%

Throughput Variance

- 334 test transfers from ANL to NERSC
- 4 types: mem-mem (84), mem-disk (78), disk-mem(87), disk-disk(85)

TABLE VI: Throughput of ANL-NERSC transfers (Mbps)

	mem-mem	mem-disk	disk-mem	disk-disk
Min	308.9	202.4	297.4	10.85
1st Qu.	1149	599.6	1265	527.3
Median	1472	819.0	1569	645.9
Mean	1463	789.6	1563	670
3rd Qu.	1772	1007	1851	841.3
Max	2706	1354	2529	1079
CV	35.69%	31.63%	30.80%	33.10%

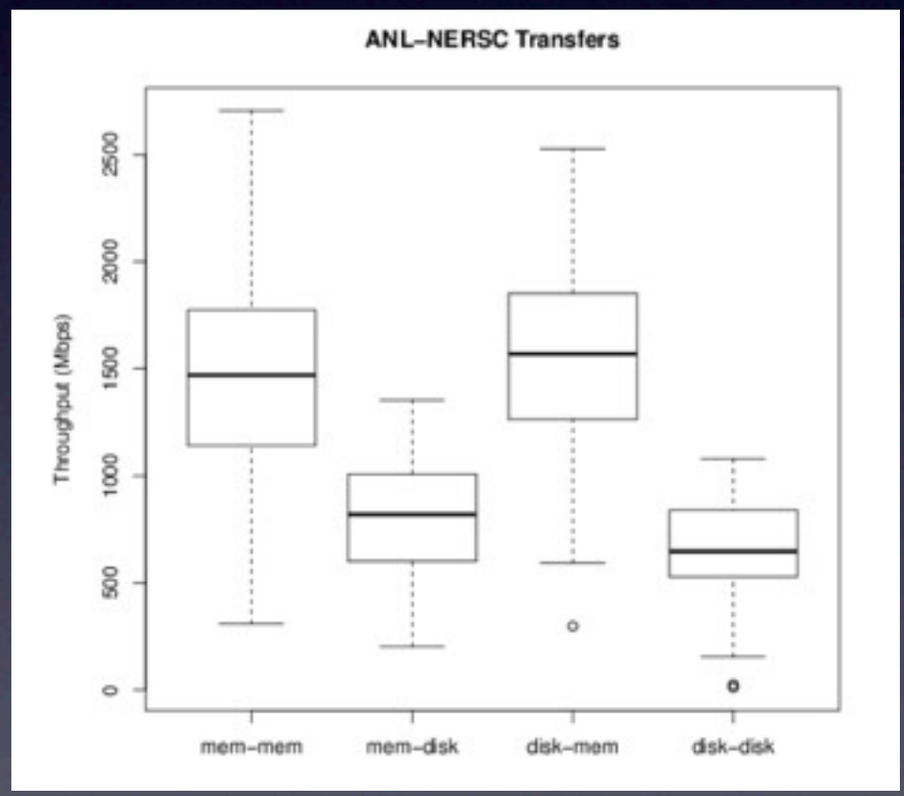


Throughput Variance

- 334 test transfers from ANL to NERSC
- 4 types: mem-mem (84), mem-disk (78), disk-mem(87), disk-disk(85)

TABLE VI: Throughput of ANL-NERSC transfers (Mbps)

	mem-mem	mem-disk	disk-mem	disk-disk
Min	308.9	202.4	297.4	10.85
1st Qu.	1149	599.6	1265	527.3
Median	1472	819.0	1569	645.9
Mean	1463	789.6	1563	670
3rd Qu.	1772	1007	1851	841.3
Max	2706	1354	2529	1079
CV	35.69%	31.63%	30.80%	33.10%



Throughput Variance

- Potential factors impacting throughput
  - GridFTP configuration
    - number of parallel TCP streams
    - number of stripes (servers)
    - reuse TCP connections (“-fast” option)
  - Competing network/server resources
    - link utilization
    - concurrent GridFTP transfers

Throughput Variance

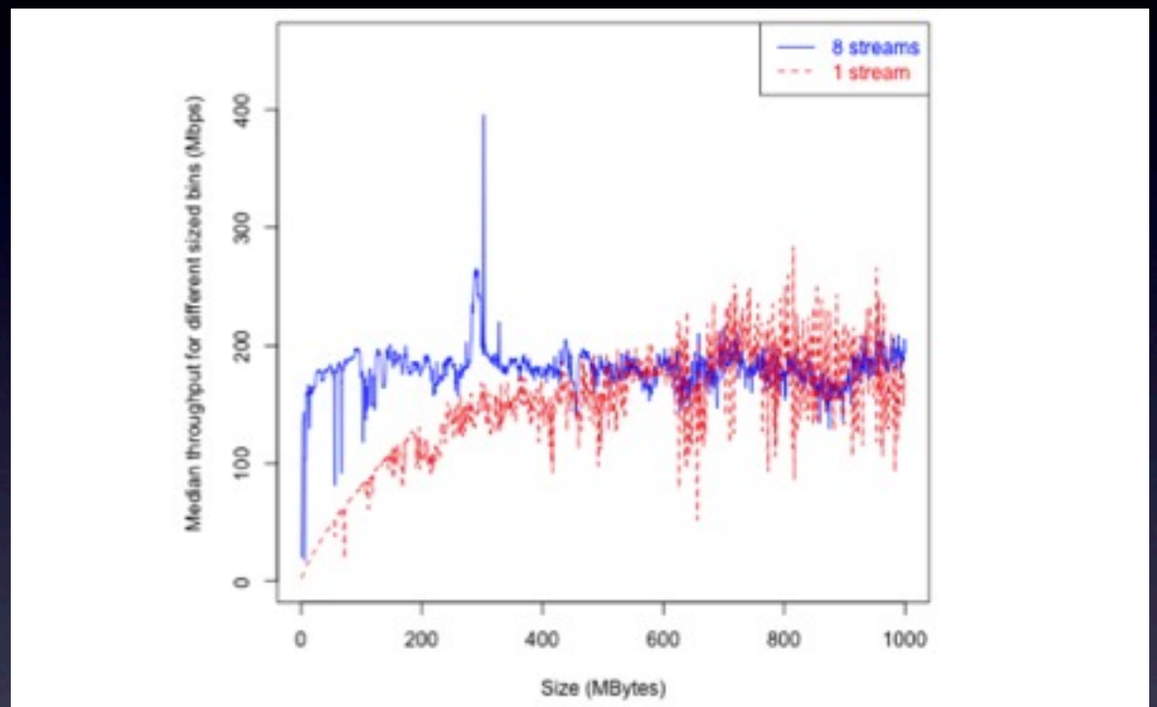


Fig. 3: Throughput of 8-stream and 1-stream transfers of size (0, 1GB)

- Analysis on SLAC-BNL data set suggests multiple TCP streams result in higher throughput for small files, possibly due to Slow Start
- For larger sizes, the throughput is roughly the same, suggesting packet losses are rare if any

Throughput Variance

- "Stripe" in GridFTP context: multiple servers participating in transfer
- Data from NCAR - NICS data set seems to suggest dependence of throughput on number of stripes

TABLE VIII: Throughput of 16GB/4GB transfers in NCAR data set (Mbps)

Year based analysis of 16GB transfers								
Year	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
2009	1076	10.79	707.3	889.3	877	1075	1543	294.17
2010	233	95.36	516.5	619.2	651.7	742.9	1150	205.53
2011	12	441.73	480.71	538.77	539.1	575.39	652.07	66.92
Year based analysis of 4GB transfers								
Year	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
2009	853	4.14	593.1	873.1	849.2	1125	1587	366.01
2010	247	72.99	767	977.1	903.1	1083	1209	225.09
2011	37	296.27	376.13	497.81	475.63	556.6	637.13	101.12

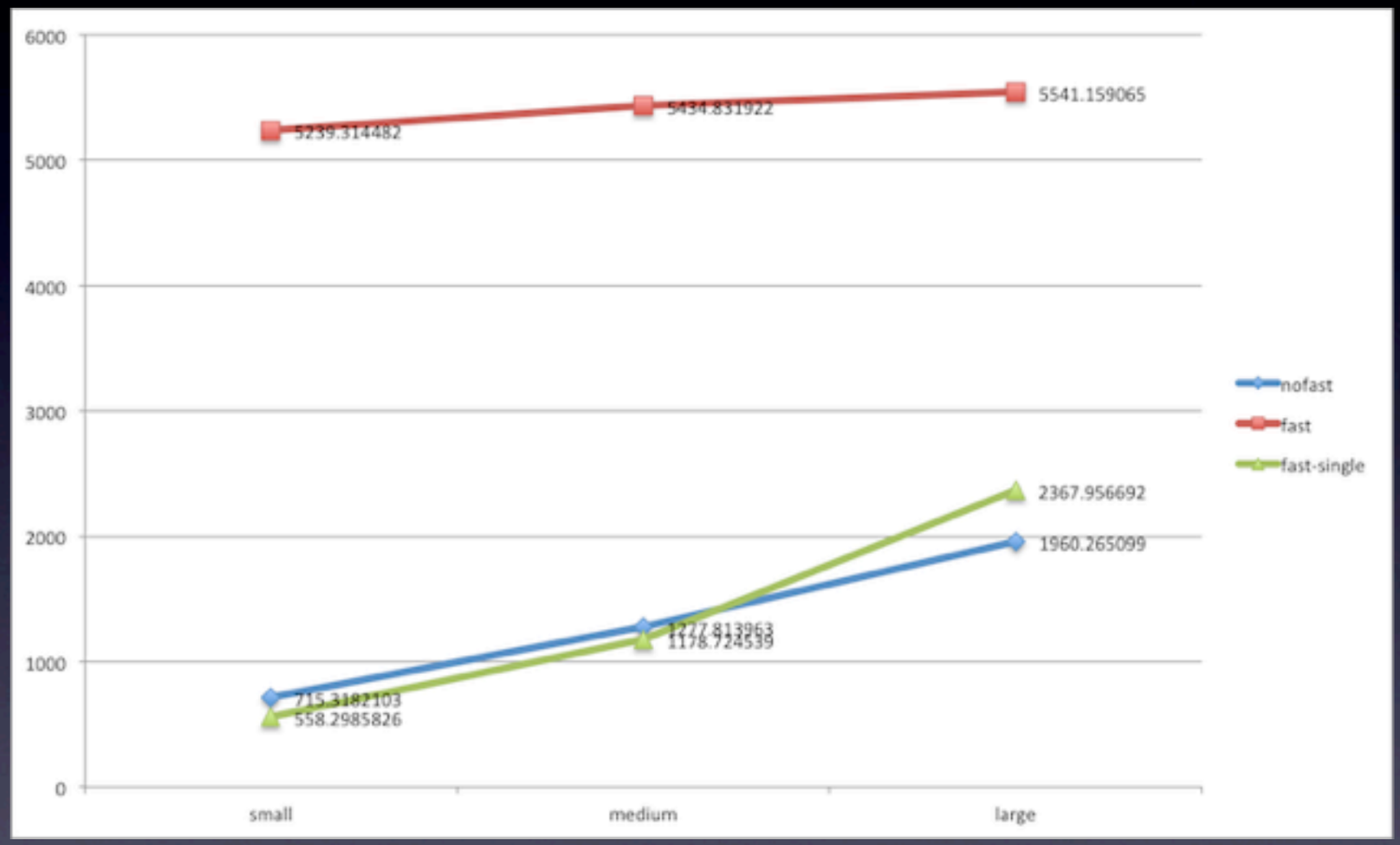
TABLE IX: Throughput of 16GB/4GB transfers in NCAR data set (Mbps)

Stripes based analysis of 16GB transfers								
No. of Stripes	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
1	13	441.73	483.7	541.75	546.84	616.48	652.07	69.88
2	547	10.79	542	714	705.4	855.2	1207	212.34
3	761	19.83	748.7	976	931.6	1150	1543	306.96
Stripes based analysis of 4GB transfers								
No. of Stripes	No. of Transfers	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Standard Deviation
1	18	372.2	449.6	506.2	569.2	574.2	1309	225.85
2	447	72.99	566.7	773.1	772.8	1021	1209	245.37
3	759	4.14	625.6	927.6	875.8	1169	1587	375.38

Throughput Variance

- GridFTP provides a parameter called "-fast", which reuses TCP connection for transfers of directories
- Since the usage log does not record the presence (or lack thereof) of this option, we conducted experiment on a testbed
- Link capacity: 10 Gbps; RTT: 48.8ms
- Test cases:
  - transferring whole directory with "-fast"
    - one process of globus-url-copy, one TCP connection
  - transferring whole directory without "-fast"
    - one process of globus-url-copy, multiple TCP connection
  - transferring whole directory with "-fast", but transfer each file one by one
    - multiple processes of globus-url-copy, multiple TCP connection

Throughput Variance



Small: 128 x 64MB
Medium: 32 x 256MB
Large: 8 * 1GB

The impact of reusing TCP connection is obvious, esp. on high BDP path

Throughput Variance

- ESnet routers provide SNMP data of byte counters at a 30s interval
- Estimate total number of bytes from SNMP counter data:

$$B_i = b_1 \frac{(\tau_{i2} - s_i)}{30} + (\sum_{j=2}^{m-2} b_j) + b_{m-1} \frac{(s_i + D_i - \tau_{i(m-1)})}{30}$$

TABLE XI: Correlation between GridFTP bytes and total number of bytes B_i (NERSC-ORNL)

	rt1	rt2	rt3	rt4	rt5
1st Qu.	0.677	0.604	0.719	0.750	0.749
2nd Qu.	0.419	0.147	0.138	0.327	0.294
3rd Qu.	0.538	0.592	0.543	0.415	0.371
4th Qu.	0.782	0.872	0.797	0.789	0.790
All	0.902	0.922	0.919	0.918	0.918

High correlations suggest GridFTP traffic are dominating

TABLE XII: Correlation between GridFTP bytes and bytes from other flows (NERSC-ORNL)

	rt1	rt2	rt3	rt4	rt5
1st Qu.	0.254	0.188	0.429	0.505	0.486
2nd Qu.	0.269	-0.067	-0.110	0.089	0.071
3rd Qu.	0.059	0.157	0.110	0.015	-0.039
4th Qu.	0.196	0.328	0.239	0.287	0.276
All	0.351	0.365	0.443	0.524	0.527

Low correlations suggest other traffic is not causing throughput variance

* SNMP data obtained with help from Chris Tracy, Jon Dugan, Eric Pouyoul, Brian Tierney and other ESnet support staffs

Throughput Variance

- Concurrent transfers

$$\tilde{t}_i = \sum_{j=1}^{j_{max}} (R - \sum_{k=1}^{n_{ij}} t_k) \times \frac{d_{ij}}{D_i} = R - \sum_{j=1}^{j_{max}} \sum_{k=1}^{n_{ij}} \frac{t_k d_{ij}}{D_i}$$

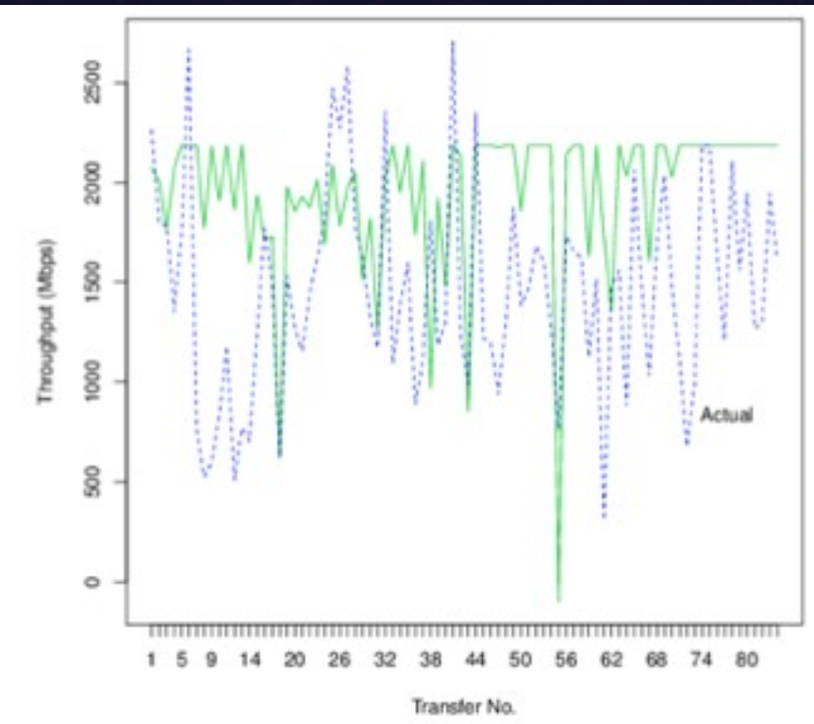


Fig. 8: Actual and predicted throughput values for memory-to-memory transfers from ANL to NERSC ($\rho = 0.2513$)

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

Tool Development for Variance Study

- We developed tools for collecting usage/performance data on servers.
  - *top*/*mpstat* measures the CPU usage
  - *tcpdump*/*tcptrace* analyzes TCP packet loss
- Controlled experiments were run on ANI LIMAN testbed to test these tools
  - developed scripts to use *double*, *tc*, and *netem* utilities to emulate CPU load and packet loss
- Deployed these tools on production NERSC and SLAC data transfer nodes (DTNs)*

* NERSC - SLAC experiments coordinated with Jason Hick and Jason Lee, NERSC, and Yee-Ting Li and Wei Yang, SLAC

NERSC - SLAC Experiments

- Between NERSC and SLAC DTNs
- Link Capacity: 10Gbps
- At every hour
  - a memory to memory transfer runs from NERSC to ANL for 30s
  - another one runs from ANL to NERSC for 30s
  - prior to the two test runs, at the 59th minute every hour, the monitoring tools we developed are run to record CPU usage data and TCP traces
- SNMP data obtained from NERSC and SLAC routers, as well as ESnet routers

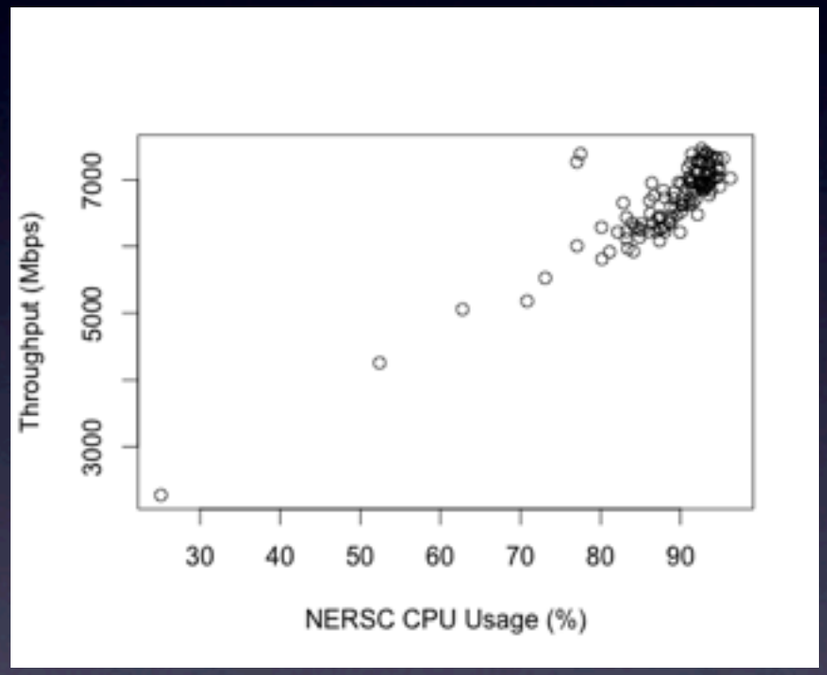
NERSC - SLAC Experiments

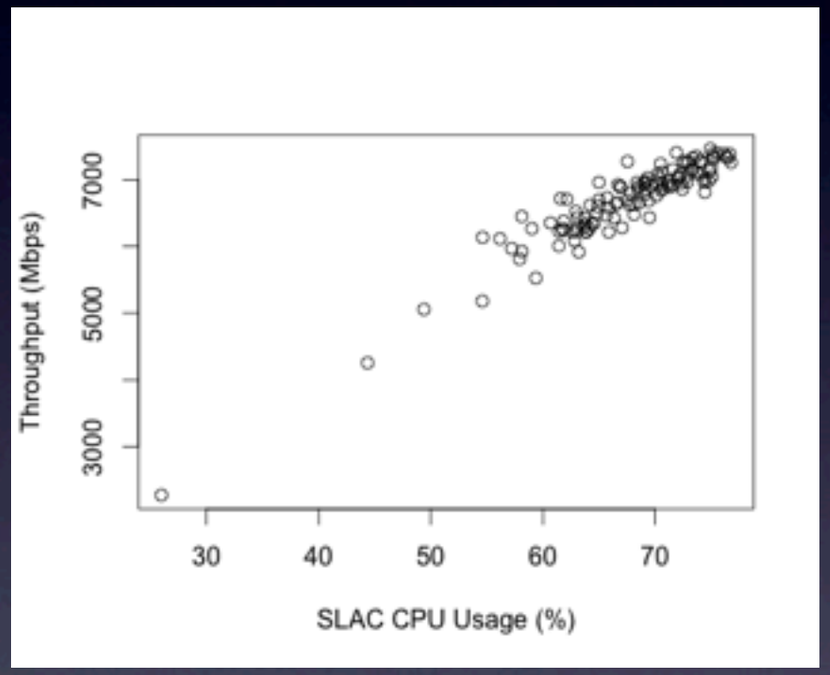
- High correlation between CPU usage and throughput

	NERSC -> SLAC	SLAC -> NERSC
Cor(Throughput, NERSC CPU Usage)	0.89	0.81
Cor(Throughput, SLAC CPU Usage)	0.94	0.90

NERSC - SLAC Experiments

- Linear Model: throughput ~ nersc_cpu + slac_cpu (NERSC -> SLAC transfers)



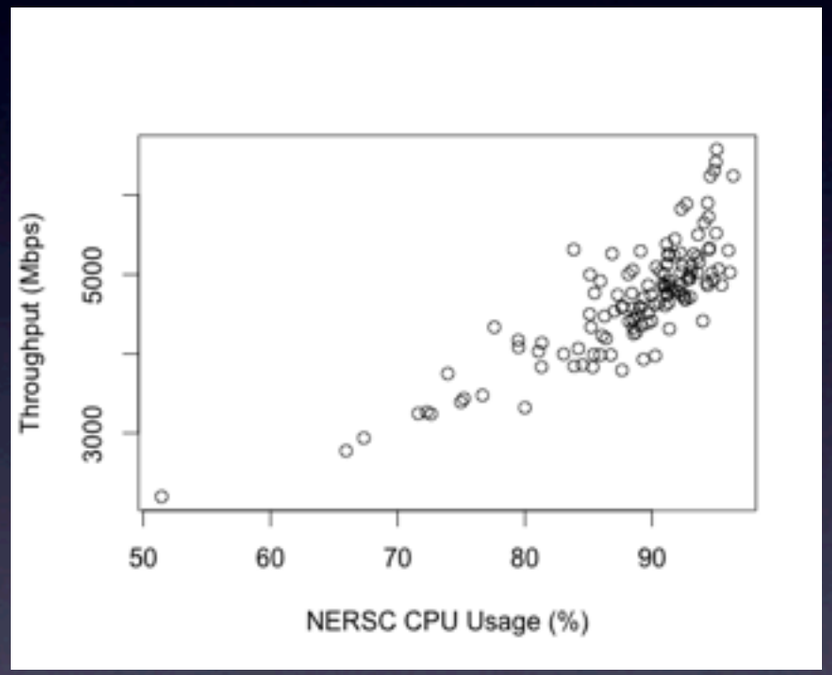


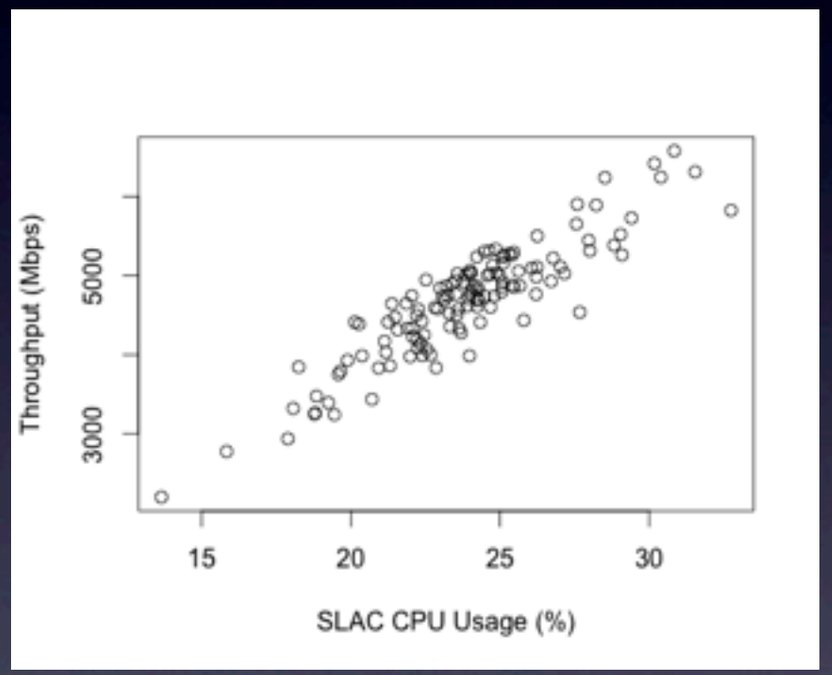
	nersc_cpu	slac_cpu	(intercept)
Estimates	29.347	58.293	155.472

Multiple R-squared: 0.9264, Adjusted R-squared: 0.9252
F-statistic: 736.8 on 2 and 117 DF, p-value: < 2.2e-16

NERSC - SLAC Experiments

- Linear Model: throughput ~ nersc_cpu + slac_cpu (SLAC -> NERSC transfers)





	nersc_cpu	slac_cpu	(intercept)
Estimates	39.203	152.655	-2427.486

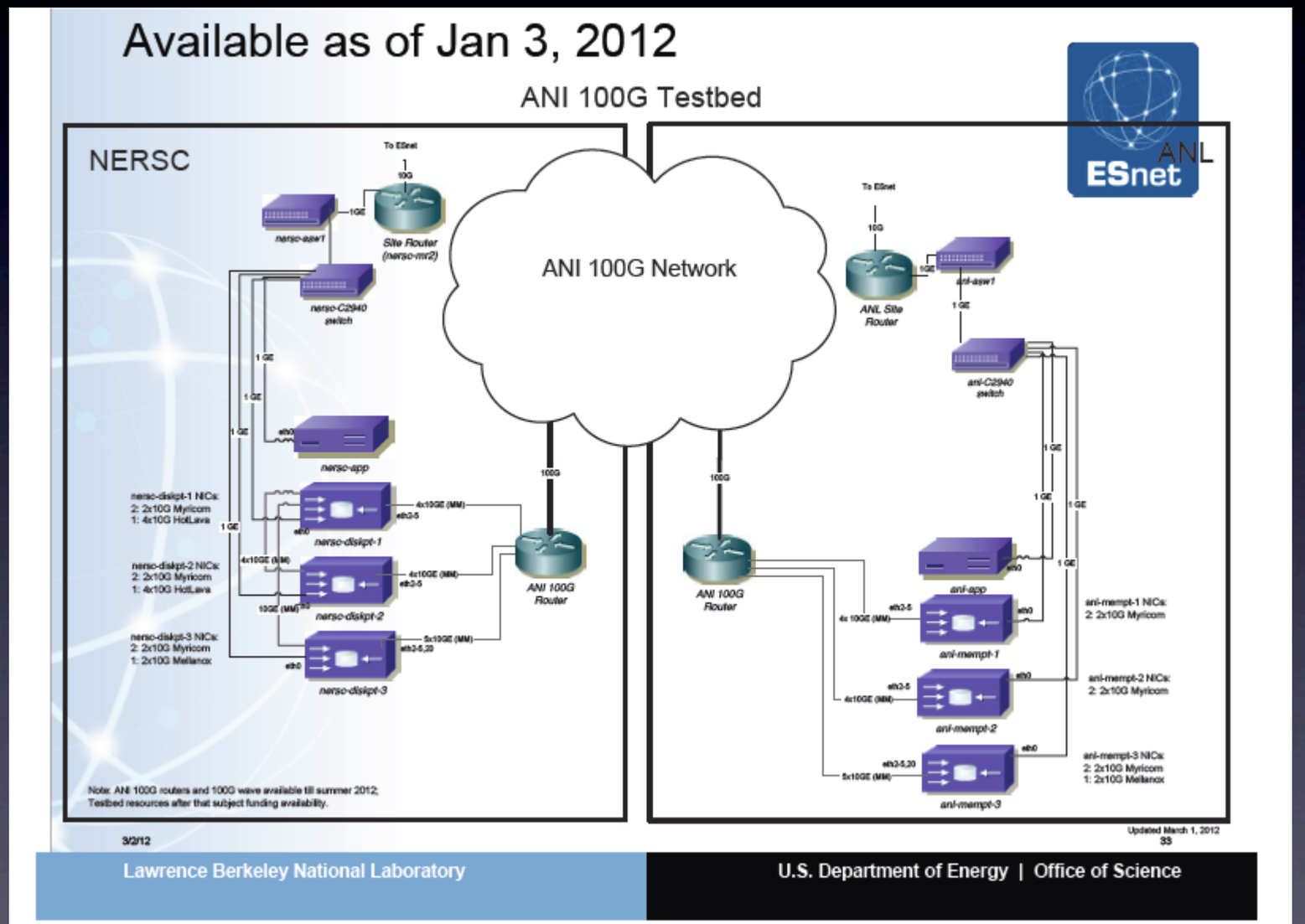
Multiple R-squared: 0.885, Adjusted R-squared: 0.8831
F-statistic: 480.8 on 2 and 125 DF, p-value: < 2.2e-16

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

DOE ANI 100G Testbed - Overview

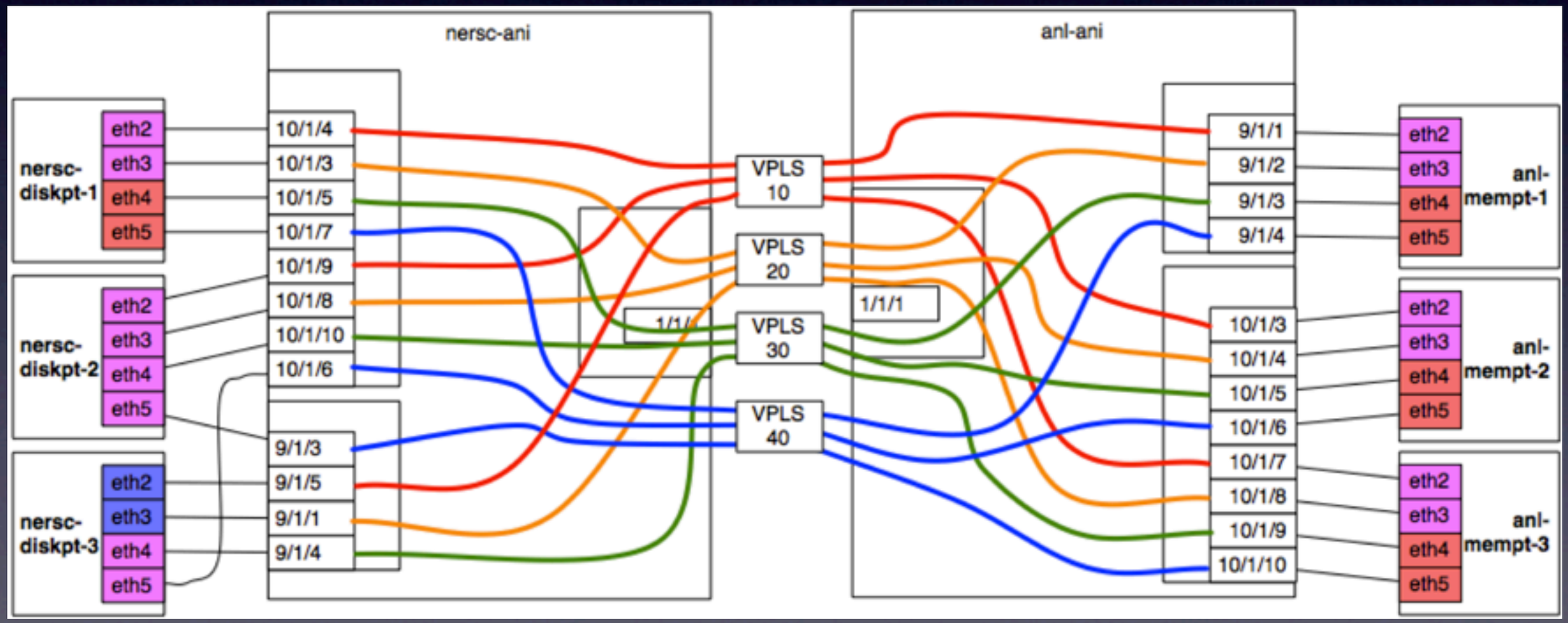
- Two sites: NERSC and ANL
- Each site has 3 performance testing hosts: "diskpt"s and "mempt"s
- Each host has 4 10 GbE interfaces



*Brian Tierney's DOE PI meeting talk, March 1-2, 2012
Work done by Eric Pouyol and Brian Tierney, ESnet

DOE ANI 100G Testbed - Overview

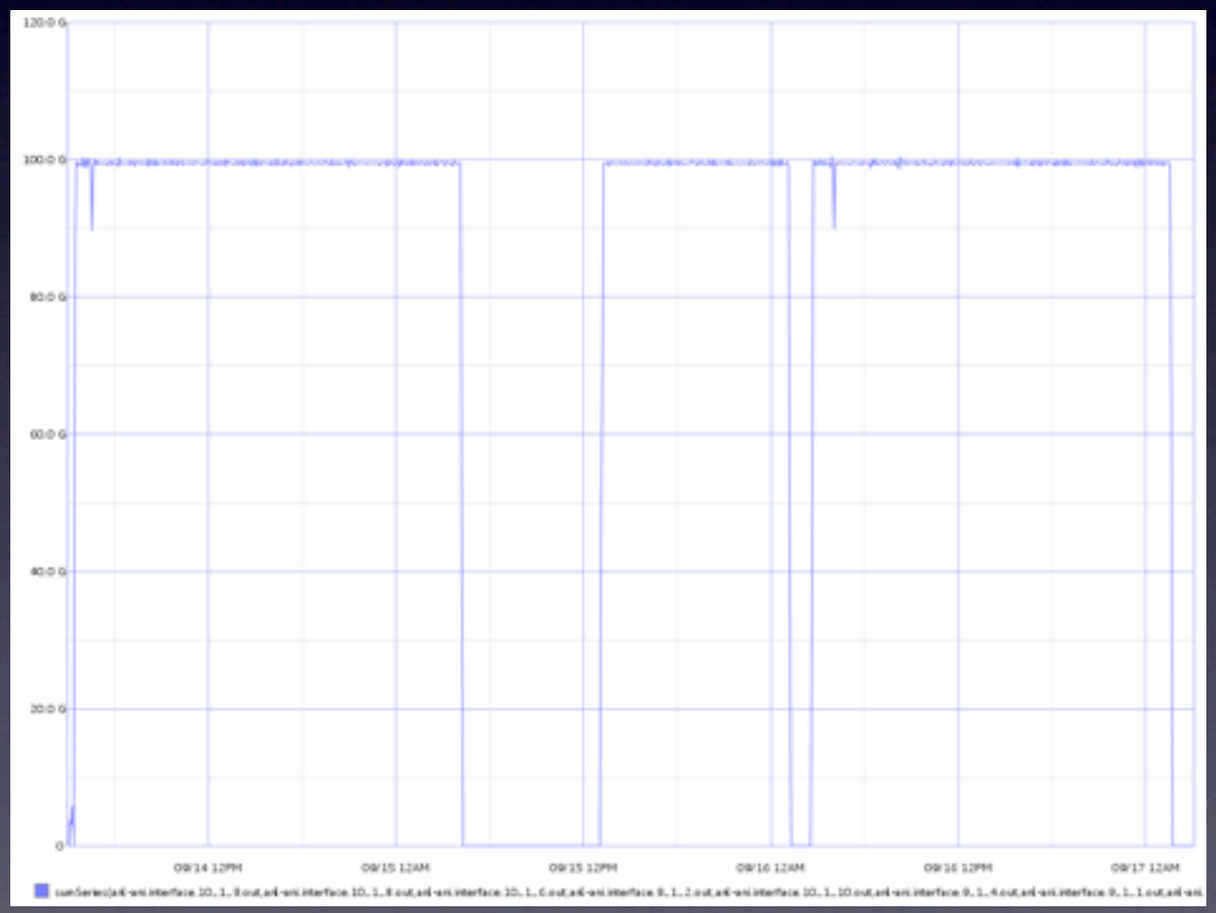
- Two sites: NERSC and ANL
- Each site has 3 performance testing hosts: "diskpt"s and "mempt"s
- Each host has 4 10 GbE interfaces

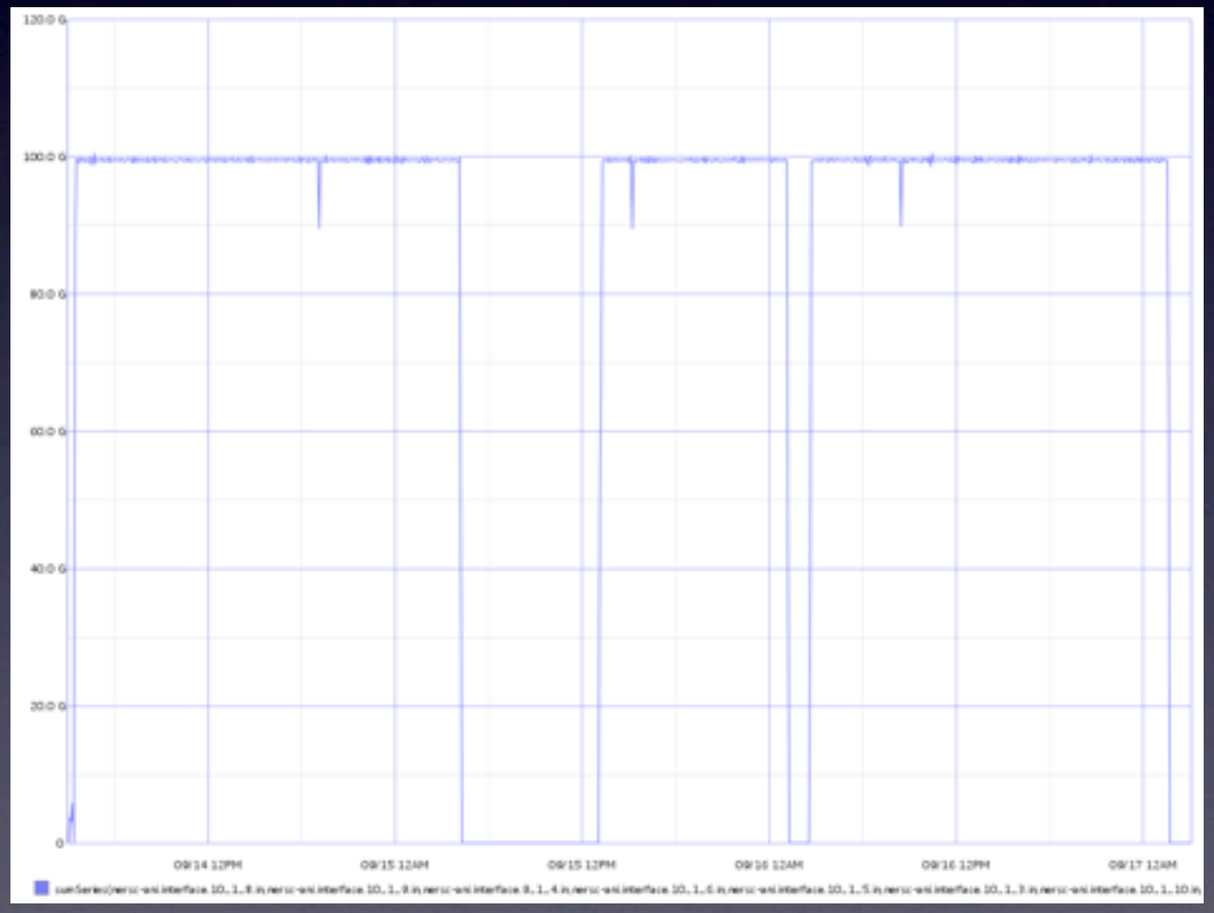


*Brian Tierney's DOE PI meeting talk, March 1-2, 2012
Work done by Eric Pouyol and Brian Tierney, ESnet

TCP Behavior on 100 Gbps Paths

- ran *nuttcp* benchmark for 20 hours; utilized all 24 interfaces on all 6 performance testing hosts
- achieved around aggregated throughput of 98.67 Gbps
- no errors reported by SNMP

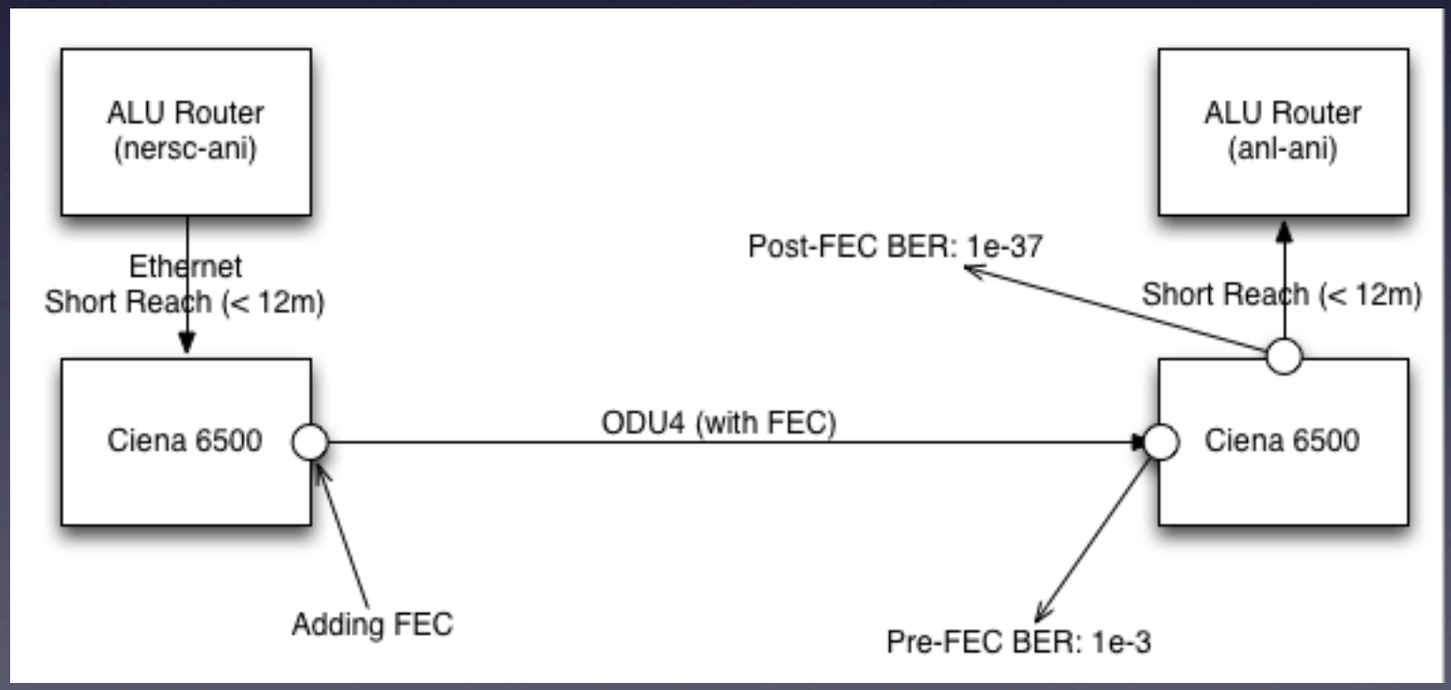




SNMP counters from NERSC and ANI 100G routers

TCP Behavior on 100 Gbps Paths

- Reason for 0 errors: Forward error correction in Ciena 6500s
- Before Ethernet frames are encapsuled in ODU4, FEC codes are added
- Pre-FEC Bit Error Rate can be as high as 1.36e-3
- Post-FEC Bit Error Rate: in the order of 1e-37
- At full 100 Gbps, we can only transfer 6.9e16 bits in 24 hours



* explained to us by Chris Tracy, ESnet

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

Outline

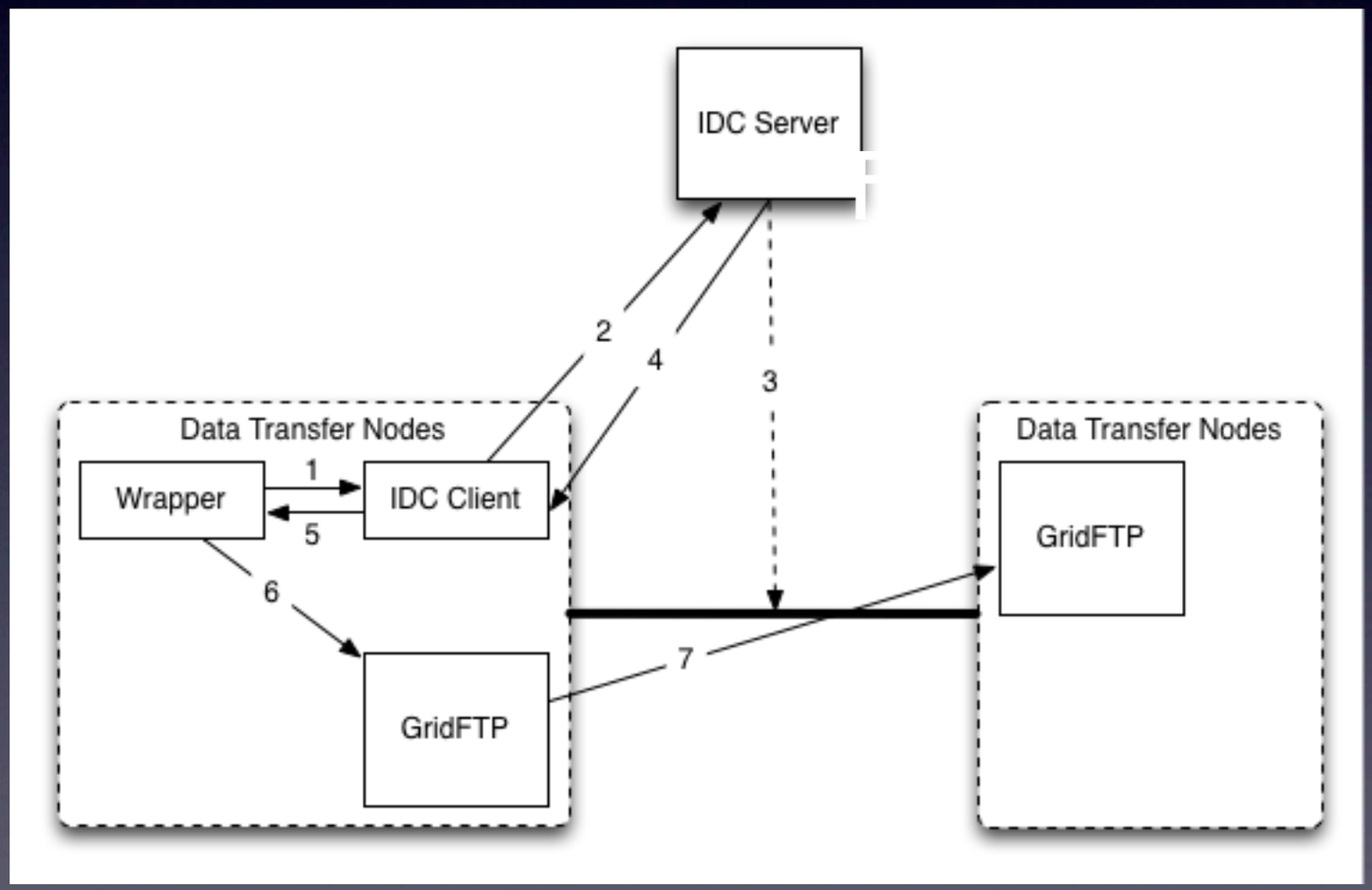
- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

GridFTP Integration with RoCE and IDC Client

- GridFTP throughput is highly dependent on available CPU resources, especially on high speed links
- RoCE has the advantage of low CPU usage; it however requires a L2 circuit
- Solution: RoCE + IDC + GridFTP

GridFTP Integration with RoCE and IDC Client

- Inter Domain Controller Protocol (IDCP) is a web-service (SOAP) based protocol for dynamic provisioning of network resources
- We are working with ESnet's implementation, OSCARS, which provides a command line interface



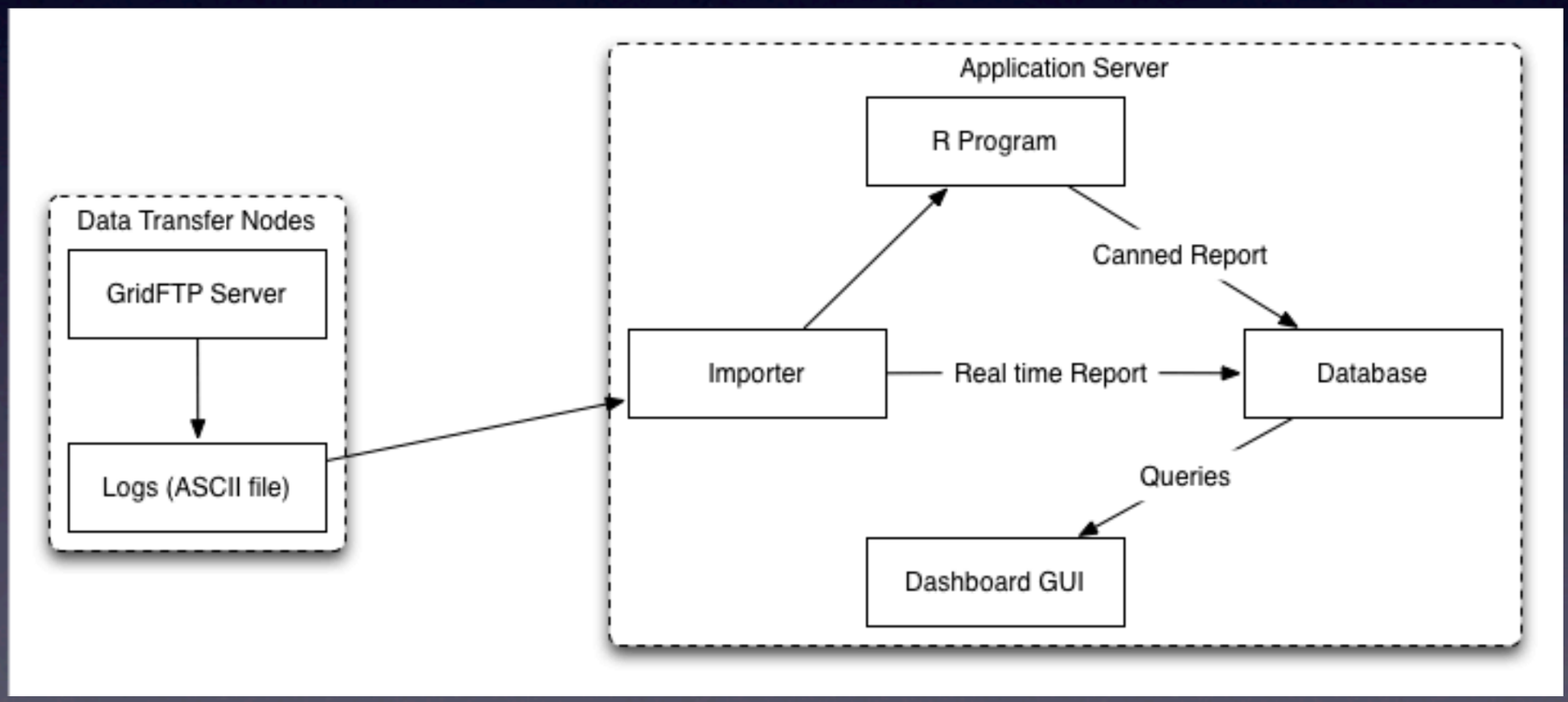
1. Wrapper calls IDC Client
2. IDC Client creates a reservation and requests a circuit to be set up
3. IDC Server talks to domain controllers and sets up the circuit
4. IDC Server notifies Client
5. Client notifies Wrapper
6. Wrapper launches GridFTP for the actual transfer
7. Transfer starts

Outline

- GridFTP Usage Log Analysis
  - Session Analysis
  - Throughput Characterization
  - Throughput Variance
- Experiments on ANI 100G and LIMAN testbeds
  - Tool Development for Variance Study
  - TCP Behavior on 100 Gbps paths
  - RoCE over L2 Circuits vs TCP over IP
- Engineering Solutions
  - GridFTP Integration with RoCE and IDC Client
  - GridFTP Dashboard GUI

GridFTP "Dashboard" GUI

- Visualizes GridFTP logs



Thank You